Summer 8-2018

Generalized Non-Inferential Approach to Modeling Restricted Discrete Choice for the Case of The Spatial Random Utility

Elena Labzina
Washington University in St Louis

WASHINGTON UNIVERSITY IN ST.LOUIS

Department of Mathematics

Generalized Non-Inferential Approach to Modeling Restricted Discrete Choice for the Case of The Spatial Random Utility

by
Elena Labzina

A thesis presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Master of Arts

August 2018
St. Louis, Missouri

Table of Contents

Acknowledgments

I am grateful to my dissertation advisers, Betsy Sinclair and Jose Figueroa-Lopez, for being a knowledgeable source of information and advice.

Summer schools and workshops I was lucky to have attended contributed a lot to my personal and academic development. I thank the organizers and participants of summer schools in University of Tampere and Central European University, EITM Institute, and Summer School in Computational Social Sciences.

Finally, I am grateful to my parents, Olya and Pavel, and my brother, Jenya, for their continuous love and support. Without them, nothing in my life could have been possible.

Elena Labzina

Washington University in Saint Louis

August 2018

Dedicated to my family.

ABSTRACT OF THE THESIS

Generalized Non-Inferential Approach to Modeling Restricted Discrete Choice for the Case of The Spatial Random Utility

by

Elena Labzina

Master of Arts in Statistics

Washington University in St. Louis, 2018

Professor Betsy Sinclair, Chair

Multinomial logistic regression model (MNL) is a powerful and easily tractable way for measuring the probabilistic impact of input variables on individual categorical choices. Crucially, the standard MNL assumes that all subjects of the study have the same choice sets. In the meanwhile, especially in political science and economics, this condition is frequently violated. Probably, the most graphical example of varying choice sets (VCS) is partially contested elections. Furthermore, the MNL implicitly implies the Independence of the Irregular Alternatives (IIA) assumption by requiring i.i.d errors that contrasts the MNL and the multinomial probit (MNP) and mixed logit (MXL) models. In the case of VCS in the MNL, the errors are correlated and IIA is clearly violated. However, neither MNP nor MXL allows estimating particular parameters for distinct choice sets. This obstacle is critical if the aim is to compare the selection process conditional on the choice restrictions. This text argues that the MNL proposes the best opportunity to model categorical choice given VCS. For that, it advances the theory of MNL adjusting this classical model for the case of VCS. Second, the paper proposes a way to calculate and evaluate the model posing minimal data restrictions. Finally, this research provides an example of the model's application.

Chapter 1

Introduction

This paper proposes a formal non-inferential generalization of the multinomial logistic (MNL) model for the case of the varying choice sets (VCS). This work reformulates this classical model making it ideologically correspondent to the current renaissance of Machine Learning (ML) making it computationally straight-forward and virtually independent of the distributional assumptions about the data.

The major contribution of the paper is the introduction and refinement of the *filter* matrix that enables neatly integrating the VCS option with the standard MNL. Another aim of this work is to provide a comprehensive mathematical formulation of all essential elements of the model. While most of the technicalities of a possible implementation are straightforward and so are not given here; code for the most sophisticated step of the estimation - the Bayesian MCMC - is provided in the Appendix. Importantly, in terms of the coefficients, the proposed model estimates the same coefficients for the spatial individual utility as the standard MNL capturing categorical choice estimates. Hence, the results from the VCS can be directly

compared to those from the classical MNL. Finally, this paper can be used as a cookbook for performing similar research.

The paper starts with the literature review presented further in this chapter. The next chapter addresses the necessary fundamentals: the notion of choice sets, the spatial random utility model, and the standard multinomial logistic regression. The major contribution of the work is presented in Chapter 3. It starts with the introduction of the notion of varying choice sets, then discusses it in the context of the independence of the irrelevant alternatives assumption, and then refines and introduces the central contribution: the filter matrix. The following chapter outlines the Bayesian approach to the estimation including the necessary priors. Also, in the second half of this chapter, it is shown that from the perspective of the problem of optimization, the proposed refinement of the MNL for VCS is equivalent to the analytically correct but computationally challenging existing approach. This work concludes by providing impressive results from the application of the introduced model to the British General elections of 2010.

Literature Review

This paper mainly speaks to two strands of literature. From the perspective of the field of Statistics, this works contributes to the research talking about the virtues of the Multinomial Logit regression model. While there are numerous papers in this group, three of them are especially important for this research. First, the text from McFadden (1973) who introduced the MNL to the scientific community in the way it is still perceived today. He was a strong proponent of this model who argued that the MNL is the best existing tool available for understanding human choice behavior. His core arguments for the MNL are that it is easily traceable, has clear distributional properties, and is computationally superior to another

popular choice model, the Multinomial probit. McFadden claimed that, above all, the MNL has excellent convergence properties and transparent estimation algorithm.

In general, the MNL is often discussed in comparison with the MNP. More than 30 years after that fundamental piece from McFadden, Dow and Endersby (2004) provide their excellent take on this comparison. They confirm McFadden's result that the MNP is more prone than the MNL to multiple estimations problems. The authors claim that, in the end, since in the best case scenario, these two models produce equivalent results, the MNL, as more robust of the two, is preferable. They rule out the most usual critique of the MNL, the independence of the irrelevant alternatives property (IIA) claiming that conditional on a particular choice setting IIA-related concerns are not relevant.

Finally, regarding its core contribution, this work in many ways is a refinement and development of Yamamoto's (2015) paper. He addresses a way to mitigate the effects of one of the potential core sources of the violation of IIA: varying choice sets (VCS). Surprisingly, there is no other work explicitly discussing the case of VCS. The likely explanation for this is that VCS is a fundamental problem only for the MNL. Other similar models, such as MLP or the multinomial mixed logit, MLX, do not require the IIA; hence their results do not become biased because of VCS. Train (2002) provides a great overview of these methods, also talking extensively about various estimation approaches.

However, Yamamoto's ideas are critical from the perspective of the electoral studies. The problem with MNP and MXL is that they do not allow to derive particular choice set-related estimates. While studying the elections, it is indeed natural to be interested in explicit comparison of the model's estimates from electoral regions that have different sets of available parties. In contrast to the MNP and MXL, the adjusted MNL allows to estimate the region-specific parameters.

Hence, logically, the second strand of literature to that this paper contributes is electoral studies. The potential of this research is not limited to this field, however, electoral studies provide, probably, the most graphical example where explicit accounting for VCS is essential. Because of the method's electoral focus, the selection of the spatial utility model as the fundamental component of the MNL is convenient (e.g Claassen et al. 2011,Galiani, Schofield, and Torrens 2014).

There are four papers that apply the core idea of Yamamoto's paper to studying elections. This main idea, an approach to adjust the MNL for VCS, is best summarized with formula (3.1) that is later discussed in detail in the text. First, his own paper looks at the municipal elections in Tokyo. Second, McAlister, Jeon, and Schofield (2015) use this formula for the Canadian elections' data. They do it without any refinement; it is still possible because Canada provides an extremely simple electoral setting: there is just one additional party in one region, Quebec. Finally, Labzina and Schofield (2015) and Labzina, Barceló, and Schofield (2017) apply the method presented in this paper. Despite that Labzina and Schofield (2015) provide a simplified overview of some contributions of this work, this text is actually a continuation of two "non-papers" earlier performed by the author. The first version of these ideas was presented in the term summary for the course "Bayesian Statistics" (Math 459) in Spring 2014. Then, in the following summer, the research was presented as a poster at Political Methodology conference 2014.

Chapter 2

Foundations

2.1 Choice Sets

Further in the text, *choice sets* refer to *discrete choice sets* available to decision makers. They can be consumers, voters, parties, firms, households, families, or any other units making decisions. Choice sets contain competing alternatives, among which the decision maker makes a *rational* choice.

A choice rule is *rational* if it satisfies the conditions of *completeness* and *transitivity* (Mas-Colell, Whinston, Green, et al. 1995). *Completeness* means that the decision maker either prefers one of each two alternatives or is indifferent between them. *Transitivity* is a mathematical property that states that if a is preferred to b and b is preferred to c, it implies that a is preferred to c. To rule out *cyclicity* of the preferences, also called *the Condorcet paradox*, *transitivity* is necessary to hold for each triple of the alternatives in the choice set.

For continuous inputs, the choice rule is usually expressed as *a utility function* that the decision maker maximizes. The next section talks about the utility function employed in this paper - *the spatial utility model*.

To be compatible with the discrete decision framework, the alternatives (and, hence, the choice rule) must satisfy three properties (Train 2009). First, the alternatives must be *mutually exclusive*, meaning that choosing one alternative implies rejecting other alternatives. Second, the set of choices must be *exhaustive*, or that all available alternatives must be included in it. This condition is another way to refer to the condition of *completeness*, which has been already mentioned. Finally, the number of the alternatives must be *finite*.

Most models, including the standard MNL, assume the *homogeneity* of the choice sets. In other words, it means that all decision makers in the sample can select from the same set of alternatives such as political parties or consumer goods. The case when some alternatives are restricted to some individuals is referred as *varying choice sets* (VCS). The paper talks about VCS in the context of the MNL in detail in the following chapter.

2.2 Individual Spatial Random Utility Model (RUM)

Decision makers select alternatives based on a utility model. If the model assumes n individuals, $i \in I$, and k alternatives, $j \in J$, then i prefers alternative s over p if $U_{is} > U_{ip}$. U_{ij} denotes the actual *unobserved* utility. The observed, modelled, utility is called *representative*: V_{is}. Since all aspects of U_{ij} are never revealed: $U_{ij} \neq \mathrm{V}_{ij}$ (Train 2009).

The typical way to perceive the relation between actual and representative utility is by adding a stochastic component: $U_{ij} = V_{ij} + \xi_{ij}$, where ξ_{ij} includes the factors not captured with the

representative utility including randomness. This is a purely schematic way of showing the relation between U_{ij} and V_{ij}: the actual relation can be more far more complex than linear.

Given the stochasticity, the decision maker makes the choice based on his or her evaluation of the probability, $P(U_{is} > U_{ip}) = P(V_{is} - V_{ip} > \xi_{ip} - \xi_{is} | s \neq p)$. Hence, the choice rule is dependent on the distributional assumptions about ξ_{ij}. Finally, choosing the best alternative z is based on the evaluation of

$$z = \underset{s}{\operatorname{argmax}} P(\cap U_{is} > U_{ij} | \forall j \in J)$$

Further this section focuses on one *spatial* way to model the *representative* utility. Then, the following section addresses one of the approaches to stochastically connect the continuous utility function to the categorical choice: MNL.

This model assumes that n individuals, $i \in I$, choose one of k alternatives, $j \in J$. Each individual has his or her preferences' location (the *bliss* point): X_i (a vector of coordinates); each alternative has its location in the same space: Z_i. The degree of the content of individual i with alternative j is based on j's proximity to i's bliss point and the valence of j (e.g Claassen et al. 2011; Galiani, Schofield, and Torrens 2014):

$v(x_i, z_j) =$ valence component$_j$ + proximity between x_i and z_j.

In case of the varying choice sets, there is also a variation in terms of the alternatives feasible to i denoted as m(i).

valence component - is the non-spatial component that characterizes the "non-spatial" attractiveness of alternative j given the available set m(i); further the linear decomposition of the valence component is assumed as $\lambda_j + \mu_{m(i),j}$

spatial component - without restricting the generality regarding the number of the dimensions the distance between x_i and z_j is further referred by the distance operator $||.||$ with the spatial factor $\beta > 0$.

If j is unfeasible for i, or $j \notin m(i)$, the utility is undefinable.

Summing up, for $\forall i, j$ the spatial individual utility:

Then:

$$v(i,j) = v(x_i, z_j | m(i)) = \begin{cases} \lambda_j + \mu_{m(i),j} - \beta||x_i - z_j|| & j \in m(i) \\ \emptyset & j \notin m(i) \end{cases} \tag{2.1}$$

2.3 Multinomial Logistic Regression (MNL)

MNL is, probably, the most popular approach to connect the utility of U_{ij} to the probability of that i selects j. If $U_{ij} = V_{ij} + \xi_{ij}$ and $y_i = j$ is i's choice: the selection of j is driven by the probability that $U_{ij} > U_{ik}$ for $\forall k \neq j \in J$. McFadden et al (1973) has shown that if the disturbances $\{\xi_i\}$ are i.i.d with the Weibull distribution:

$$F(\xi_{ij}) = exp(-e^{\xi_{ij}})$$

then

$$P(y_i = j) = \frac{exp(v_{ij})}{\sum_{\forall k \in J} exp(v_{ij})} \tag{2.2}$$

Importantly, in the end, this result is based on the distribution of the differences between $\{\xi_i\}$'s that have the Type I or extreme distribution. This distribution is close to the normal distribution and most researchers agree that in applications it can be assumed normal (Dow and Endersby 2004). Hence, often the explicit assumption of MNL is that $\{\xi_i\}$'s are normally i.i.d.

The main advantage of MNL is its tractability and computational feasibility. Also, as McFadden et al (1973) states, MNL has superior convergence preferences, especially, compared to another popular model, MNP, which is prone to have hardly detectable flat areas around the equilibrium.

A crucial consequence (and a potential weakness) of this model's distributional properties is the implicit implication of the *Independence of the Irrelevant Alternatives* (IIA) condition that is going to be addressed in the following chapter (Dow and Endersby 2004).

Chapter 3

MNL with Varying Choice Sets (VCS): Formal Model

3.1 Why is VCS problematic?

As already said in the previous section, categorical selection models including the MNL, assume the homogeneity of the choice sets. Formally, it means that if m(i) denotes the alternatives available to i and J is the full choice set, then $m(i) = J,\ \forall i \in J$. Meanwhile, in political science and economics, variation of m(i) is not rare. In political science, one example of such situations is partially contested elections typical on the local municipal level, when some parties may not run in every district (Yamamoto 2014). In Canada (Gallego et al. 2014), the UK (Labzina and Schofield 2015), and Spain (Labzina, Barceló, and Schofield 2017), not all parties compete in every region during the General elections. In economics, consumer goods may be not available on all markets. Further, the paper assumes that a significant portion of the existing alternatives is available in all possible types of distribution. Hence it

is legitimate to assume the number of all existing distributions is not greater than the full number of the alternatives, k.

For the further discussion, let's stick to the linear stricture: $U_{ij} = X_{ij}\alpha_j + \xi_{ij}$ (in the vector form), to which the spatial utility corresponds.

While MNP and MXL allow for the correlation of the residuals, hence formally they are applicable, the specification of these models does not let track the differences between the choice sets (Dow and Endersby 2004). For example, this means that they do not allow to estimate the region-specific parameters in the setting of partially contested general elections.

Meanwhile, the MNL provides an opportunity to specify the model to reflect the properties of particular choice sets. However, the classical formula 2.2 would lead to the violation of the i.i.d assumption about the errors. This would happen because the residuals for the observations such that m(i) = m(j) = $M \subset J$ are correlated, reflecting the restriction imposed by M.

As a solution to this problem, the analytic adjustment proposed by Yamamoto (2015) is to explicitly account for the feasibility of the alternatives:

$$P(y_i = j) = \begin{cases} \frac{exp(v_{ij})}{\sum_{\forall j \in m(i)} exp(v_{ik}))} & j \in m(i) \\ 0 & j \notin m(i) \end{cases} \tag{3.1}$$

(3.1) fixes the feasibility mispecification implied by (2.2): now only individually feasible alternatives are assumed to have positive probability of the selection.

Unfortunately, introducing (3.1) is not enough to get actual estimates of the model. First, even if formally the potential correlation between the errors is removed, we still may be concerned about the feasibility of the normal i.i.d assumption of the errors. The following chapter will talk about ways to mitigate these concerns by employing the Bayesian approach to the computation.

Second, despite being theoretically precise, (3.1) is computationally problematic, since it is unclear how to implement the summation for each m(i) during the estimation of the likelihood function. Further in the text, the paper shows how to construct a computationally feasible formula for the MNL with VCS. Furthermore, in the following chapter, the newly proposed formula and (3.1) are proven to be computationally equivalent by showing that they produce equivalent log-likelihoods

3.2 The IIA Assumption

Before proceeding to the major contribution, it is necessary to address the IIA assumption, since it is often mentioned as one of the critical features of the MNL model. Why is the ability to explicitly account for VCS is essential even if the interest is the estimates for the universally available options?

To begin, *the Independence of the Irrelevant Alternatives* assumption (IIA) is a fundamental concept from the decision theory that is implicitly embedded in the MNL by the i.i.d condition on the residuals. The idea of the IIA is that the binary choice between each two alternatives is independent of other alternatives and their feasibility. Mathematically, according to Austen-Smith and Banks 1999, the classical textbook about Social Choice and Decision Theory, for a choice set, J:

Definition A binary choice rule φ is IIA if for $\forall\ i \in I$, $\forall J^{'}, J^{''} \subset J$, and $\forall j_1, j_2 \in J^{'}, J^{''}$ $\varphi_i(J^{'}, \{j_1, j_2\}) = \varphi_i(J^{''}, \{j_1, j_2\})$

In the case of the standard MNL, it is always the case that $J^{'} = J^{''} = J$. Hence, despite that it is impossible to claim that the IIA holds for all possible subsets of J, conditional on the universal feasibility of all alternatives from J, the IIA formally holds. Substantively, this means that if the model always deals with the full set of J, what happens with the preferences if some alternatives are unfeasible simply lies out of the problem's domain. Hence, the IIA is satisfied by default. However, what happens in the MNL model with the VCS, in other words, if $\exists i$ for which the preferences are observed only on $J_i \neq J$? On the surface, this is not a problem since the interest is in the average estimates. Since these i never selects the alternatives outside of J_i, these unfeasible alternatives are implicitly implied to have zero utility for i. Hence, the standard MNL, even in the presence of the VCS, must provide the valid estimates. However, this conclusion is incorrect.

To begin with, the explicit sign that something is wrong with this approach (the standard MNL in the case of the VCS) is the possibility to predict non-zero probabilities for individually unfeasible alternatives lying outside J_i. This is the consequence of the violation of the IIA, which must hold for the estimation to be correct. Without the VCS adjustment, the model assumes that all i's have equal conditional preferences, $\varphi_i = \varphi$. However, in case of VCS, for various i's the choice is observed on various subsets of J.

In formal terms, the following situation takes place. The standard MNL model assumes that for $\forall i, j$: $\varphi_i = \varphi_j$. Let's assume that the choice sets $J_i \neq J_j$ such that $J_i \cup \{k\} = J_j$. Then, $\forall m \in J_i\ \varphi_i(m, k) = \varphi_j(m, k)$ must hold. In the meanwhile, $\varphi_i(m, k) = \emptyset$, or is out of the domain of the problem, since $k \notin J_i$, and $\varphi_j(m, k) \neq \emptyset$ by definition since $k \in J_j$. Hence, $\varphi_i(m, k) \neq \varphi_j(m, k)$ and the IIA is formally violated, and assuming the IIA and $J_i = J_j = J$

always produces biased estimates for the predicted probabilities. Importantly, since the predicted probabilities are constrained ($\forall i : \sum_{\forall j} p_{ij} = 1$), the estimates must end up biased even for the universally feasible alternatives.

Let's show that the adjustment proposed in the previous section and summarized with (3.1) fixes the above-described sign of the violation of the IIA. This is true since the variations in $\{\varphi_i\}$ are allowed by (3.1), and, according to this formula, the probability to select an unfeasible alternative is zero.

3.3 The *Filter* Matrix

The major problem with the formula (3.1) is that it is computationally challenging to account for the inclusion of an alternative in m(i). Also, the approach proposed so far does not allow to evaluate any common properties of the family, {m(i)}.

To fix these problems, this paper introduces the *filter* matrix, n>>k:

$$\Phi(\{i\}) = \begin{pmatrix} m(1) \\ m(2) \\ ... \\ m(n) \end{pmatrix} = \begin{pmatrix} \phi_{11} \ \phi_{12} \ ... \ \phi_{1k} \\ \phi_{21} \ \phi_{22} \ ... \ \phi_{2k} \\ ... \\ \phi_{n1} \ \phi_{n2} \ ... \ \phi_{nk} \end{pmatrix}, \ \phi_{ij} = \begin{cases} 1 & j \in m(i) \\ 0 & j \notin m(i) \end{cases} \tag{3.2}$$

Then, the number of linearly independent columns equals the number of the types of the choice bundles that must be the same as $rank(\Phi)$. For example, if there are k alternatives and for some individuals alternative k is not available:

$$rank \begin{pmatrix} 1 & 1 & ... & 1 \\ 1 & 1 & ... & 1 \\ & ... & & \\ 1 & & ... & 1 \\ 1 & & ... & 0 \end{pmatrix} = rank \begin{pmatrix} 1 & 1 & ... & 1 \\ 1 & 1 & ... & 0 \end{pmatrix} = 2 \tag{3.3}$$

Then (2.2) given (3.2) can be transformed:

$$P(y_i = j) = \frac{exp(u(i,j))\,\phi_{ij}}{\sum_{\forall k \in J} exp(u(i,k))\phi_{ik}} \tag{3.4}$$

This matrix is denoted *filter matrix* because it "filters out" of unfeasible alternatives. In particular, $u(i,j) = \emptyset \implies \phi_{ij} = 0 \implies P(i,j) = 0$. Then, (3.4) applies for $\forall i, j$ and makes the computation straight-forward. Meanwhile, for the standard MNL, where $\forall i, j\ \phi_{ij} = 1$: (3.4) = (2.2). Hence, it becomes a partial case of the proposed extension of 3.4.

3.4 Refinement of the Filter Matrix

The structure of Φ is clearly redundant and can be simplified since Φ is likely to contain a lot of equal rows that correspond to the same choice bundles. Let's denote the modification of Φ that contains only distinct rows that represent different choice bundles as $\Phi^{'}$. Then, trivially, $rank(\Phi) = rank(\Phi^{'})$.

Also, let's introduce the inversible operator $M : M(\Phi) \longmapsto \Phi^{'}$ fully described by the column-vector $M^{'} \in R^n$, that maps Φ and $\Phi^{'}$:

$$\phi^{'}_{m^{'}(i),j} = \phi_{i,j} \tag{3.5}$$

As assumed before, the number of all possible distributions is not greater than k. Then, instead of nk distinct values this method deals at most with $n + k^2$ values, which decreases the number of the input variables by at least $k(n - k - 1)$. In context of large datasets, the change in the volume of the stored information is approximated by $\lim_{n\to\infty} \frac{nk}{n+k^2}$.

By L'Hospital's Rule (Taylor 1952),

$$\lim_{n\to\infty} \frac{f(x)}{g(x)} = \lim_{n\to\infty} \frac{f^{'}(x)}{g^{'}(x)},$$

which implies:

$$\lim_{n\to\infty} \frac{nk}{n + k^2} = \lim_{n\to\infty} \frac{k}{1} = k$$

Illustrative example

Hence, the proposed adjustment results in k times less stored information.

To provide a tentative example of how Φ can be transformed into a combination of $M^{'}$ and $\Phi^{'}$:

$$\Phi = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix} \implies \Phi^{'} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}, M^{'} = \begin{pmatrix} 1 \\ 2 \\ 2 \\ 1 \\ 1 \\ 3 \\ \end{pmatrix} \tag{3.6}$$

Hence, the number of the stored variables decreases from 18 to 15. However, for n = 100,000,000 and the same $\Phi^{'}$ the economy will be more impressive: the decrease is from 300,000,000 to 100,000,003.

What is the intuition behind $\Phi^{'}$ and $M^{'}$? $\Phi^{'}$ defines *distinct* types of choice sets existing in a setting and implicitly assigns each of them an order number corresponding to the row number where it is described. $M^{'}$ maps each observation in the sample to a distinct choice set specified in $\Phi^{'}$.

An illustrative substantive example is an electoral setting where different sets of parties run in country regions (Labzina and Schofield 2015, Labzina, Barceló, and Schofield 2017). $\Phi^{'}$ has as many rows as there are the regions and $M^{'}$ just denotes the region of each individual.

Summing up this section, the generalized logistic link for $\forall i, j$:

$$P(y_i = j) = \frac{exp(u(i,j))\, \phi^{'}_{m_i^{'}j}}{\sum_{\forall k \in J} exp(u(i,k)) \phi^{'}_{m_i^{'}k}} \tag{3.7}$$

Chapter 4

Estimation of the Model

4.1 Bayesian Approach as an Non-Inferential Method

As said before, the standard MNL way to account for uncertainty may be problematic because of likely violations of the normal i.i.d assumption in the case of VCS. As said in the very beginning, the work proposes to use the non-inferential approach, meaning significantly relaxing data distributional assumptions.

In the inferential approach, the error is an additive part of the latent utility function. For example:

$$u_{ij} = v_{ij} + \epsilon_{ij},\ \epsilon_j \sim EV(0,1) \tag{4.1}$$

Meanwhile, in the Bayesian approach the additive error term is not included in the linear input of the model. Then, how does the model account for the uncertainty?

The general form of the utility model this paper estimates (2.1) contains $1 + (J - 1) + (\sum_{\forall m(i)} |m(i)| - 1)$ unknown parameters: $\theta = (\beta, \{\lambda_j\}^1, \{\mu_{m(i),j}\}^2)$.

While in the standard inferential statistics these parameters would be considered unknown and constant, the Bayesian philosophy assumes them random variables. The Bayesian approach methods allow to derive the distributions of the parameters that provide the best fit for the existing data.

Formally, the *posterior* distribution of the parameters, $p(\theta|y)$, is obtained based on a *prior* distribution, $p(\theta)$, and the likelihood of the data conditional on the parameters, $L(y|\theta)$:

$$p(\theta|y) \propto p(\theta)L(y|\theta) \tag{4.2}$$

While in some cases (relatively rare in a real setting) the posterior can be mathematically derived, the advantage of the Bayesian approach is the existence of iterative procedures that simulate the posterior distributions without finding the analytic solution. Bayesian Markov Chain Monte-Carlo (MCMC) methods, as, for example, the three-step Metropolis-within-Gibbs algorithm from Yamamoto 2014, work well even with non-informative priors. As the canonical textbook on Bayesian statistics (Gelman et al. 2014) states, the convergence to the proper posterior given a sufficient number of the iterations is not dependent on the priors.

The empirical independent of the final solution to the specification of the prior is the key to why this method can be claimed as *non-inferential*. Despite the necessity to select the priors, certain uninformative priors are known to provide the converge of the model.

[1] $\lambda_1 = 0$ as in the standard MNL
[2] $\mu_{m(1),1} = 0$

Currently multiple packages exist that compute the posterior distribution of the parameters via MCMC. One of them, *JAGS* with its R library *rjags*, is provided as the example of implementation in the appendix. While being flexible regarding the data, the implementation still requires an explicit specification of the prior description of the parameters and the likelihood function of the data given the parameters.

4.1.1 Priors

The following non-informative priors are based on Yamamoto 2014 and have been successfully empirically tested in Labzina and Schofield 2015 and Labzina, Barceló, and Schofield 2017:

$$\beta \sim N(0, 1/1000) \tag{4.3}$$

$$\lambda_j \sim N(0, \tau_\lambda)\,,\ \tau_\lambda \sim Gamma(1/10, 1/10),\ \forall j > 0 \tag{4.4}$$

$$\mu_{m(i),j} \sim N(0, \tau_\mu)\,,\ \tau_\mu \sim Gamma(1/10, 1/10),\ j \in m(i)\ \forall\, m(i) \tag{4.5}$$

4.2 Likelihood Function for the MNL with VCS

Let's show that the newly proposed formula for probability (3.7) produces the same log-likelihood as the original analytically valid formula (3.1), first proposed by Yamamoto (2015), and so, these formulas are equivalent from the perspective of the computational optimization.

First of all, let's look at the log-likelihood for the standard MNL explicitly derived from the probability function (2.2):

$$L_0(.) = \sum_{i=1}^{n}(\sum_{j=1}^{J} I(y_i = j)u(i,j) - log(\sum_{j=0}^{J} e^{u(i,j)})) \tag{4.6}$$

Meanwhile, for the newly proposed formula for MNL with VCS (3.7), the log-likelihood is

$$L_1(.) = \sum_{i=1}^{n}(\sum_{j=1}^{J} I(y_i = j)(u(i,j) + log(\phi^{'}_{m_i^{'}j})) - log(\sum_{j=0}^{J} e^{u(i,j)\phi^{'}_{m_i^{'}j}})) \tag{4.7}$$

Formula (4.7) looks different compared to the log-likelihood for the fundamentally correct (3.1):

$$L_2(.) = \sum_{i=1}^{n}(\sum_{j=1}^{J} I(y_i = j)u(i,j) - log(\sum_{j \in m(i)} e^{u(i,j)})) \tag{4.8}$$

Let's prove that $L_1(.)$ is equivalent to $L_2(.)$

Since $I(y_i = j) = 1 \implies j \in m(i) \implies log(\phi^{'}_{m_i^{'}j}) = log(1) = 0$ and $I(y_i = j) = 0$ zeroes out undefined in that case $log(\phi^{'}_{m_i^{'}j})$. Hence, (4.7) simplifies to:

$$\sum_{i=1}^{n}(\sum_{j=1}^{J} I(y_i = j)u(i,j) - log(\sum_{j=0}^{J} e^{u(i,j)}\phi^{'}_{m_i^{'}j})) \tag{4.9}$$

Then,

$$\sum_{i=1}^{n}(\sum_{j=1}^{J} I(y_i = j)u(i,j) - log(\sum_{j \in m(i)} e^{u(i,j)\phi'_{m'_i j}} + \sum_{j \notin m(i)} e^{u(i,j)\phi'_{m'_i j}})) =$$

$\sum_{i=1}^{n}(\sum_{j=1}^{J} I(y_i = j)u(i,j) - log(\sum_{j \in m(i)} e^{u(i,j)(1)} + \sum_{j \notin m(i)} e^{u(i,j)(0)})) =$

$\sum_{i=1}^{n}(\sum_{j=1}^{J} I(y_i = j)u(i,j) - log(\sum_{j \in m(i)} e^{u(i,j)} + \sum_{j \notin m(i)} 1)) =$

$\sum_{i=1}^{n}(\sum_{j=1}^{J} I(y_i = j)u(i,j) - log(\sum_{j \in m(i)} e^{u(i,j)}) + constant,$

which is equivalent to

$$\sum_{i=1}^{n}(\sum_{j=1}^{J} I(y_i = j)u(i,j) - log(\sum_{j \in m(i)} e^{u(i,j)})) = L_2(.)$$

Hence, it has been proven that $L_1(.)$ and $L_2(.)$ are equivalent.

Chapter 5

Application of the Model: British General Elections in 2010

The rest of the paper provides an actual example of the method's use. It is based on Labzina and Schofield (2015). This section applies the MNL with VCS to the evaluation of the British General elections.

5.1 Model

The UK has five major parties: Labour (1), Conservatives (2), Liberal Democrats (3), Scottish National Party (SNP) (4), and Plaid Cymru (PC) (5). Labour, Conservatives in LibDems run in all three British electoral regions: England (1), Scotland (2), and Wales (3). SNP runs only in Scotland. PC runs only in Wales:

$$\Phi^{'} = \begin{pmatrix} 1\ 1 & 1\ 0 & 0 \\ 1\ 1 & 1\ 1 & 0 \\ 1\ 1 & 1\ 0 & 1 \end{pmatrix} \tag{5.1}$$

The spatial model:

$$u(x_i, z_j) = \lambda_j - \beta||x_i - z_j|| + \mu_{jr(i)} + \epsilon_i \tag{5.2}$$

Meanwhile, because of the structure of (5.1), several assumptions are made to simplify the model. Only μ_{12}, μ_{13}, μ_{22}, and μ_{23} are estimated; for all other combinations of j and r: $\mu_{jr} = 0$.

The main aim of this analysis is to evaluate the mixed party *valences*. Because of the model's specification, the mixed party valences are defined as:

$$\lambda'_j = \lambda_j + \frac{1}{n} \sum_{r(i):\mu_{jr} \neq 0} n_r \mu_{jr(i)}, \tag{5.3}$$

where r(i) denotes the region of individual i.

Summing up:

$$\lambda'_1 = \lambda_1 + \frac{1}{n}(\mu_{12} n_2 + \mu_{13} n_3)$$

$$\lambda'_2 = \lambda_2 + \frac{1}{n}(\mu_{22} n_2 + \mu_{23} n_3)$$

$$\lambda'_3 = 0$$

$$\lambda'_4 = \lambda_4$$

$$\lambda'_5 = \lambda_5$$

Finally, assuming no correlation between the terms, the conservative way to estimate the standard error:

$$sd(\lambda'_j) = \sqrt{Var(\lambda_j) + \frac{1}{n^2} \sum_{r(i):E(\mu_{jr})\neq 0} n_r^2 Var(\mu_{jr})} \tag{5.4}$$

Also, it is necessary to be able to predict the probabilities, so the voting predictions can be compared to the actual results. First, since according to the logit model, the individual voting likelihoods are predicted as $p_{i\,j} = \frac{e^{v_{i\,j}}}{\sum\limits_{k\in r(i)} e^{v_{i\,k}}}$, where r(i) is the region of an individual, the voting prediction for j for the region r(i):

$$\hat{p_{j\,r}} = \frac{\sum\limits_{i\in r} p_{i\,j}}{n_r} \tag{5.5}$$

Then for the whole country, the predictions are:

$$\hat{p_j} = \frac{1}{n} \sum_{r\in r(i)} n_r \, \hat{p_{j\,r}}$$

Assuming the zero correlation of the probabilities across the regions, the conservative estimates of the standard errors for the probabilities on the national level:

$$sd_{p_j} = \sqrt{Var(\,p_j\,)} = \sqrt{\frac{1}{n^2} \sum_{r\in m(i)} n_r^2 Var(\,p_{j\,r}\,)}$$

5.2 Data

This section provides a brief description of the data. For a more detailed overview, see the main article with the results (Labzina and Schofield 2015). To do the estimation, the model requires micro-level data that contains the *dependent* categorical scalar variable (the party voted) and the independent vector variable (the individual position in the ideological two-dimensional space). Also, the model utilizes the categorical variable indicating the region of an observation (England, Wales, or Scotland).

British Electoral Study 2009-2010[3] provides all this necessary information. The used sample contains 8084 observations, with 777 observations from the individuals from Scotland and 370 individuals from Wales. The individuals from Northern Ireland are excluded from the study at the very beginning, since this region has an entirely distinct set of political parties compared to other regions; the total of 8084 does not include them.

Regarding the dependent variable: 3,097 individuals in the sample reported to have voted for the Conservatives, 2,350 - for the Labour party, 2,384 - for the Liberal Democrats, 210 - for the Scottish National Party (SNP), and 43 - for Plaid Cymru (PC) – the party running only in Wales.

The independent vector variable – the individual ideological position – is estimated using the component factor analysis (CFA) based on a number of representative survey questions. The aim of this procedure is to reduce the number of the ideological dimensions getting two individual-level variables that are weakly correlated. In the case of the U.K., the two dimensions are *nationalism* (the attitude towards the E.U.) and *economy* (left-right). 5.1 shows the estimated distribution of the ideology for sample. The list of the survey question

[3] The British Election Study at The University of Essex, http://bes2009-10.org/

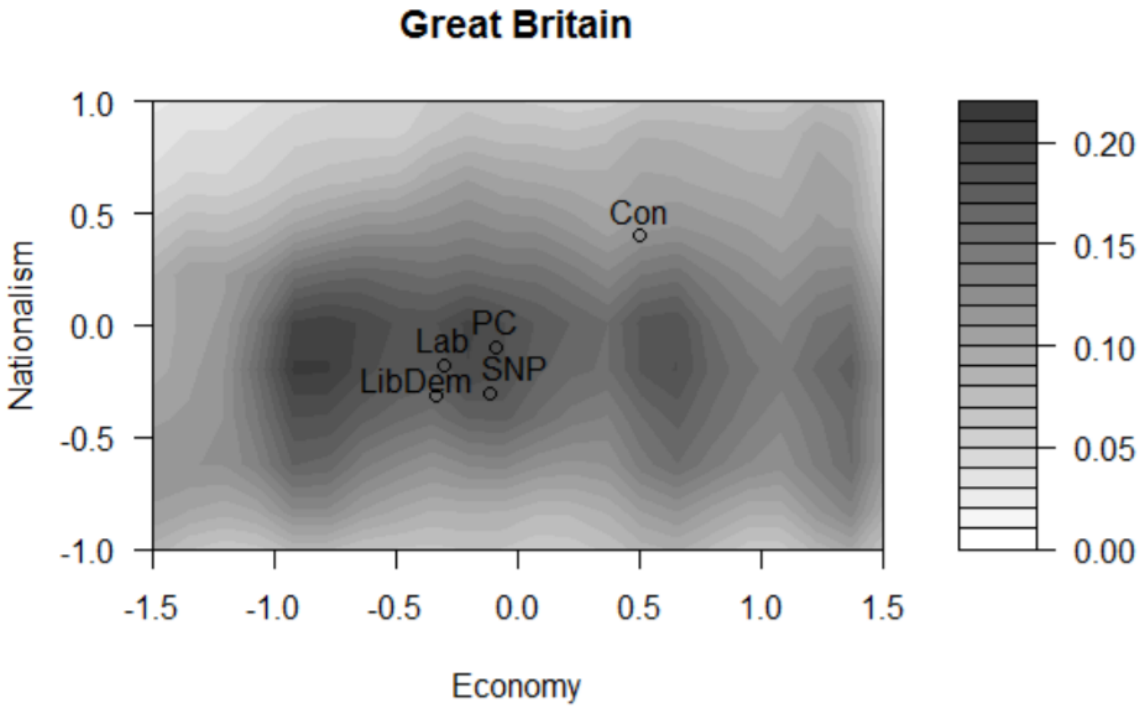

Figure 5.1: CFA: Distribution of the two ideological components in the sample (from Labzina and Schofield 2015

and a detailed description of the employed CFA procedure can be found in Schofield, Gallego, and Jeon 2011.

5.3 Results

Based on the non-informative priors introduced in the beginning of this chapter, the Bayesian MCMC model is specified (n = 8084). It runs 3 chains of 25,000 iterations with R package rjags. The model has a confident convergence according to the *Gelman-Rubin* and *Heidelberger-Welch* diagnostics (Gelman et al. 2014). Table 5.1 shows that the estimated model almost perfectly predicts the actual vote shares in the sample.

Table 5.1: Predicted and sample voting probabilities

	England			Scotland		
	sample	est	conf interval (95%)	sample	est	conf interval (95%)
p_{Lab}	0.280	0.280	[0.277, 0.283]	0.354	0.352	[0.346, 0.359]
p_{Con}	0.414	0.413	[0.407, 0.420]	0.167	0.167	[0.156, 0.182]
p_{LibDem}	0.307	0.307	[0.303, 0.310]	0.216	0.216	[0.211, 0.221]
p_{SNP}	-	-	-	0.262	0.263	[0.258, 0.267]
p_{PC}	-	-	-	-	-	-
	Wales			All sample		
p_{Lab}	0.355	0.352	[0.339, 0.365]	0.291	0.291	[0.291, 0.291]
p_{Con}	0.285	0.286	[0.261, 0.311]	0.383	0.383	[0.382, 0.384]
p_{LibDem}	0.250	0.251	[0.239, 0.262]	0.295	0.295	[0.294, 0.295]
p_{SNP}	-	-	-	0.026	0.026	[0.026, 0.026]
p_{PC}	0.111	0.112	[0.109, 0.114]	0.005	0.005	[0.005, 0.005]

Table 5.2 presents the estimations of the coefficients from the model. Importantly, only μ_{Con3}'s credible interval contains 0, all other estimates have confident direction of their effect.

Table 5.3 presents the regional and mixed valences. Expectedly, valences differ across the regions.

Counterfactual analysis

So far the model has shown an excellent performance: the estimates are significant and correspond to the substantive considerations about the British politics. Hence, the last question to ask: what would happen if the standard MNL were applied to the same data? How would the valence coefficients relate to the mixed valencies from the adjusted MNL?

Table 5.4 presents the results from the counterfactual MNL. Table 5.5 contrasts the estimates from the standard MNL and the MNL with VCS with their calculated[4] confidence/credible intervals. As expected, the MNL highly underestimates the valences of PC and SNP.

[4]To estimate the confidence/intervals, the standard formula for the confidence interval, $\bar{x} - 1.96 * \sigma/\sqrt{n}, \bar{x} - 1.96 * \sigma/\sqrt{n}$, was applied to the results from Table 5.2, 5.3, and 5,4

Table 5.2: Estimation results

	estimate	credible interval (95 %)
β	0.873	[0.833, 0.913]
λ_{Lab}	-0.102	[-0.163,-0.039]
λ_{Con}	0.259	[0.193, 0.325]
λ_{SNP}	0.227	[0.024,0.432]
λ_{PC}	-0.762	[-1.127,-0.409]
μ_{Lab2}	0.589	[0.391,0.792]
μ_{Con2}	-0.466	[-0.729,-0.208]
μ_{Lab3}	0.458	[0.193, 0.726]
μ_{Con3}	-0.056	[-0.363,0.247]
DIC	Mean/penalized deviance 15175/15184	
N	8084	

Table 5.3: Regional mixed valence

	England	Scotland	Wales	Mixed
λ_{PC}	-	-	-0.762	-0.762
λ_{Lab}	-0.102	0.487	0.091	0.002
λ_{Con}	0.259	-0.470	0.622	0.213
λ_{SNP}	-	0.227	-	0.227

Furthermore, the coefficients for β and λ_{con} are wrongly estimated as well and this distinction is statistically significant. For λ_{lab}, the MNL with VCL cannot differentiate the estimated coefficient from zero while the usual MNL's result is that λ_{lab} is significantly negative. Overall, the standard MNL tends to underestimate the spatial effect captured with β and overestimate the party effects captured with other coefficients.

Table 5.4: Counterfactual MNL

base=LibDem	
Variable	Est
	(|t-stat|)
β	0.761***
	0.017
λ_{Lab}	-0.024
	0.029
λ_{Con}	0.281***
	0.031
λ_{SNP}	-3.903***
	0.084
λ_{PC}	-5.489**
	0.160
n	8084
LL	-8396.6
McFadden R^2	0.141

Table 5.5: Comparison of the results: MNL with VCS (mixed valence) (1) versus MNL (2) with derived 95%-confidence/credible intervals

	(1)	(2)
β	0.873	0.761
	0.833, 0.913	0.761, 0.761
λ_{lab}	0.002	-0.024
	-0.060, 0.064	-0.025, -0.023
λ_{con}	0.213	0.281
	0.146, 0.279	0.280, 0.282
λ_{snp}	0.227	-3.903
	0.024, 0.432	-3.905, -3.901
λ_{pc}	0.762	-5.489
	1.127, 0.409	-5.492, -5.485

Bibliography

Austen-Smith, David, and Jeffrey S Banks. 1999. *Positive political theory*. Vol. 1. University of Michigan Press.

Claassen, Christopher, et al. 2011. "Estimating the effects of activists in two-party and multi-party systems: comparing the United States and Israel." *Social choice and welfare* 36 (3-4): 483–518.

Dow, Jay K, and James W Endersby. 2004. "Multinomial probit and multinomial logit: a comparison of choice models for voting research." *Electoral studies* 23 (1): 107–122.

Galiani, Sebastian, Norman Schofield, and Gustavo Torrens. 2014. "Factor Endowments, Democracy, and Trade Policy Divergence." *Journal of Public Economic Theory* 16 (1): 119–156.

Gallego, Maria, et al. 2014. "The variable choice set logit model applied to the 2004 Canadian election." *Public Choice* 158 (3-4): 427–463.

Gelman, Andrew, et al. 2014. *Bayesian data analysis*. Vol. 2. CRC press Boca Raton, FL.

Labzina, Elena, Joan Barceló, and Norman Schofield. 2017. "Valence and Ideological Proximity in the Rise of Nationalist Parties: Spanish General Elections, 2008 and 2011." In *State, Institutions and Democracy*, 105–142. Springer.

Labzina, Elena, and Norman Schofield. 2015. "Application of the variable choice logit model to the British general election of 2010." In *The Political Economy of Governance*, 313–333. Springer.

Mas-Colell, Andreu, Michael Dennis Whinston, Jerry R Green, et al. 1995. *Microeconomic theory*. Vol. 1. Oxford university press New York.

Schofield, Norman, Maria Gallego, and JeeSeon Jeon. 2011. "Leaders, voters and activists in the elections in Great Britain 2005 and 2010." *Electoral Studies* 30 (3): 484–496.

Taylor, Angus E. 1952. "L'Hospital's rule." *The American Mathematical Monthly* 59 (1): 20–24.

Train, Kenneth E. 2009. *Discrete choice methods with simulation*. Cambridge university press.

Yamamoto, Teppei. 2014. *A multinomial response model for varying choice sets, with application to partially contested multiparty elections*. Tech. rep.

Appendix

Input parameters

- Y - the dependent categorical variable (n x 1) $\in \{1, 2, ...J\}$
- $D = ||X - Z||$ - the distance between the position of the individual and each alternative (n x J) $\in \mathbb{R}$, for those pairs where $j \notin m(i)$ any value d_{ij} is acceptable
- M - the number of the type of the individual choice bundle (n x 1) $\in$ [1,..., rank(Φ)]
- F - the description of the types of the individual choice bundles (rank(Φ) x J) $\in \{0, 1\}$

Typically, Y is already provided in the data, D can be explicitly calculated from the data, and M and F require additional encoding based on the research assumptions.

Bayesian simulation of one estimation of the generalised MNL: JAGS

This subsection provides the code for the Bayesian estimation of the generalized MNL for the JAGS-file, that is called from R via the library *rjags*. Without restricting the generality, the example is given for the priors provided in 2.3 and the tentative filter matrix from 2.2.1.

```
GMNL_jags = function()
{
    for(i in 1:N)
    {
     for(k in 1:K)
     {
        v[i,k] <- lambda[k] + mu[m[i],j] + beta * D[i]
```

```
    expv[i,k] <- exp(v[i,k])*phi[m[i],k]
    pv[i,k] <- expv[i,k]/sum(expv[i,1:K])
  }
  y[i] ~ dcat(pv [i, 1:K])
}

beta ~dnorm(0,1/1000)

lambda[1] <- 0
lambda[2] ~ dnorm(0, taul)
lambda[3] ~ dnorm(0, taul)

for (t in 1:3)
{
    for (p in 1:3):
    {
        mu[t,p] ~ dnorm(0, taun)
    }
}

mu[1,1] <- 0

taul ~ dgamma(.1,.1)
taun ~ dgamma(.1,.1)

phi[1,1] <- 1
```

```
    phi[1,2] <- 1
    phi[1,3] <- 1
    phi[2,1] <- 1
    phi[2,2] <- 0
    phi[2,3] <- 1
    phi[3,1] <- 0
    phi[3,2] <- 1
    phi[3,3] <- 1
}
```